あらゆる AI ワークロードの基盤となる AI Hypercomputer の最新アップデート！

Google Cloud Next Tokyo

関本 信太郎

Google Cloud
カスタマー エンジニア

片岡 義雅

Google Cloud
カスタマー エンジニア

Agenda

01 AI Hypercomputer 最新アップデート

02 AI インフラ運用の難しさ

03 まとめ

01. AI Hypercomputer 最新アップデート

AI is in our DNA

AI 関連の研究開発

- 数千人規模のリサーチャー
- 7k + の論文
- Transformer, OCS

最先端の基盤モデル

- Gemini
- PaLM 2
- Imagen
- LaMDA
- GLaM
- BERT

グローバル規模のアプリ運用経験

堅牢でオープンなSW エコシステム

TensorFlow

PyTorch

Kubernetes

#1 OSS contributor

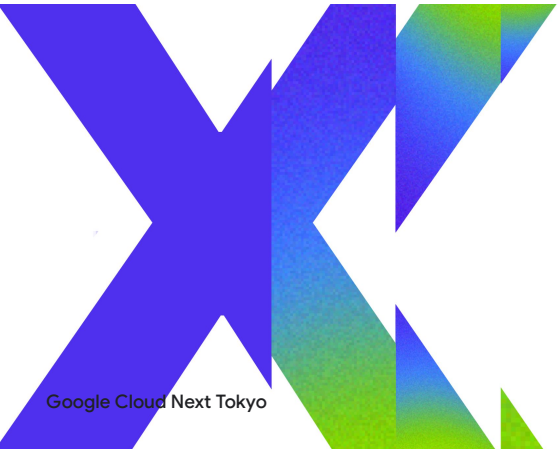
Source: OSS Index

世界規模のインフラストラクチャ

- 目的に沿って設計されたコンピュート
- ストレージ
- ネットワーク
- コンテナ
- サステナビリティ

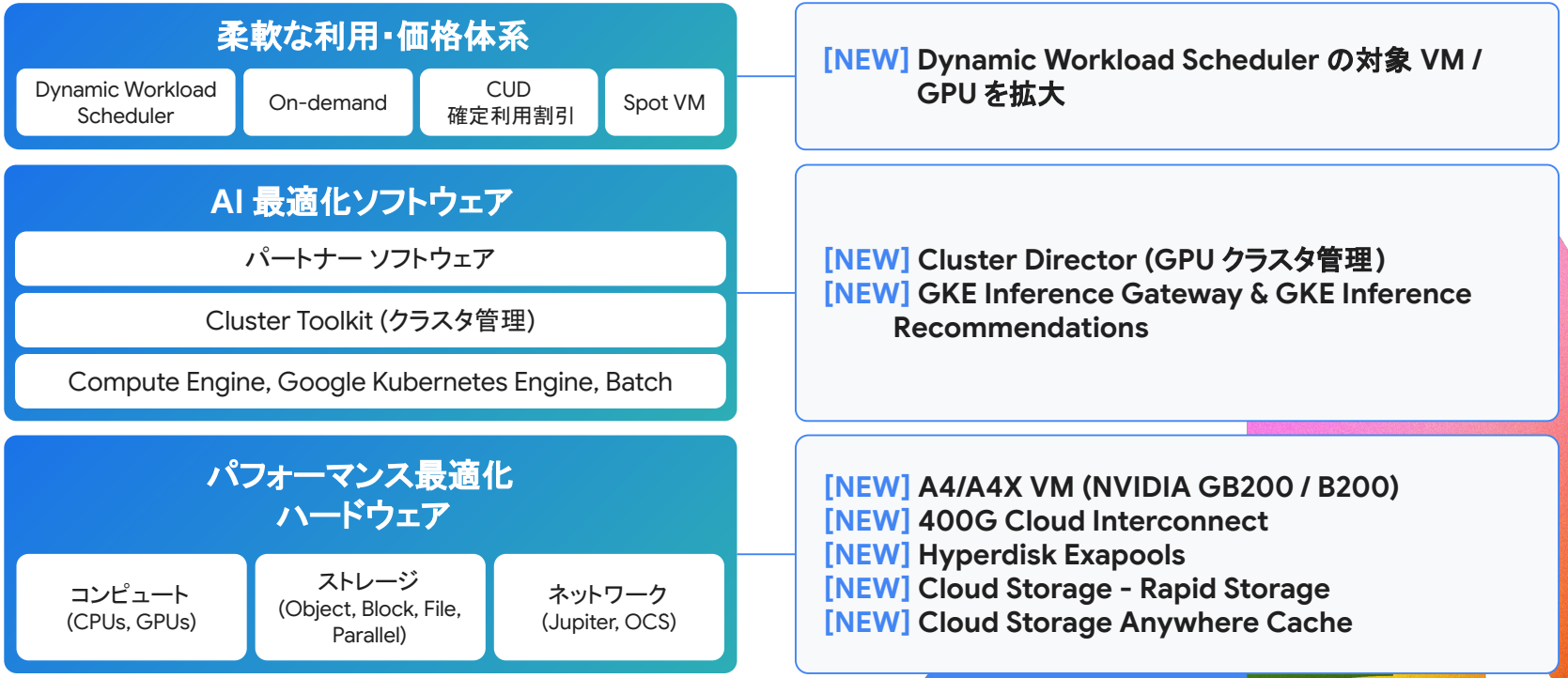
AI Hypercomputer

- 最先端のアクセラレータが選択可能
- 高いパフォーマンスとコスト効率
- 大規模クラスタでの信頼性とセキュリティ
- すぐに始められて柔軟にスケール可能



AI Hypercomputer のアーキテクチャ

AI Hypercomputer のアップデート



最新の GPU のポートフォリオ

A3 Ultra (H200)

さまざまな用途の AI ワークロード

3.2 Tbps の GPU 間通信

一般提供開始

A4 (B200)

大規模な AI モデルの学習と推論

3.2 Tbps の GPU 間通信

2.2x の計算性能 (H200 比)

一般提供開始

New

A4X (GB200 NVL72)

Reasoning モデルに最適化

28.8 Tbps の GPU 間通信

Google の先進的な液冷方式

プレビュー

New

RTX Pro 6000

推論やデジタルツイン

AI グラフィックスやオムニバース

小規模モデルの推論

プレビュー

Titanium ML network adapter

NVIDIA ConnectX-7 RoCE NIC integrated into our network fabric

Cluster Director

TPU の進化

2015 2018 2020 2022 2023 2024

v1

1x/chip
Inference

インターナルの推論用アクセラレータ

v2

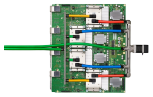
1x/chip
1x/pod

分散共有メモリ搭載

v3

3x/chip
12x/pod

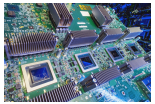
液冷の導入



v4

6.6x/chip
100x/pod

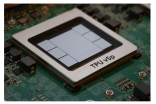
チップ間通信に OCSを用いることで柔軟に制御可能



v5e

4x/chip
Inference

大規模学習や推論のコスト効率を最適化



v5p

21x/chip
750x/pod

最も柔軟なアクセラレータ

v6e (Trillium)

100x performance

次世代フロンティア AI モデルの発掘

※ パフォーマンス比較は対v2 比

Ironwood

2025 年末にリリース予定の第 7 世代 TPU

5X チップあたりのピーク パフォーマンス

6X チップあたりの HBM 容量

2X 電力効率

* vs Trillium
JAX や PyTorch に最適化

AI アクセラレータのまとめ

*G4D & Ironwood TPU は予定

GPU TPU

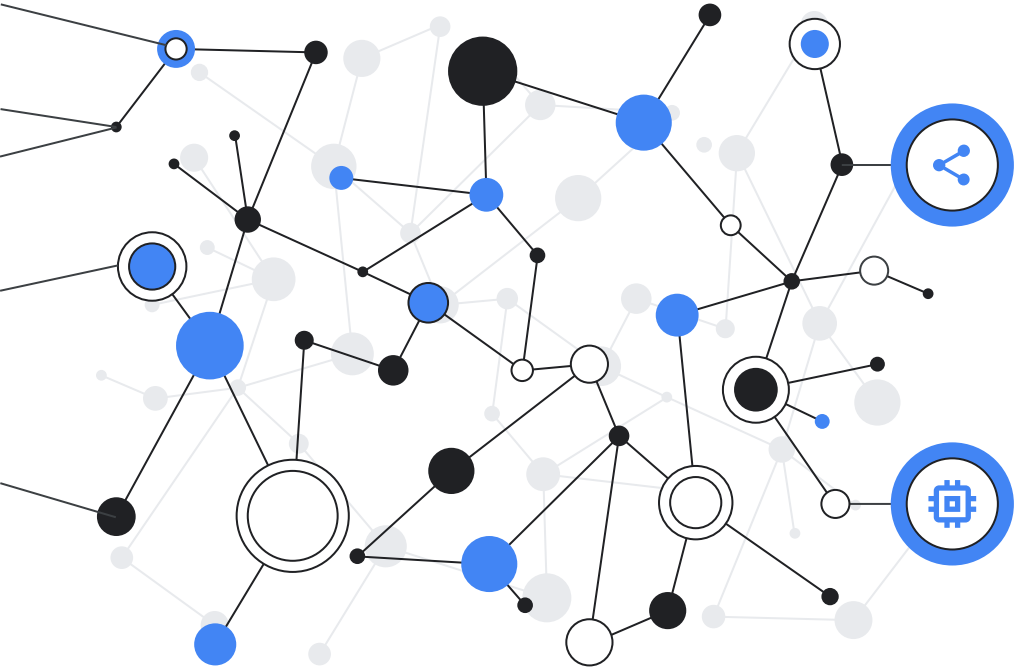
AI インフラとしての GKE

Kubernetes による学習・推論基盤

Capabilities include:

- 高速で高度なスケジューリング
- プラネタリースケールの分散処理
- アプリ基盤で使い慣れた
  インターフェースでの運用管理

巨大なスケールのサポート



シームレスな AI コンピュート基盤を支える

65,000

ノード / クラスター

マルチスライス設定により

50,000

一つのモデルで利用できるチップ数

Multi Cluster Orchestrator

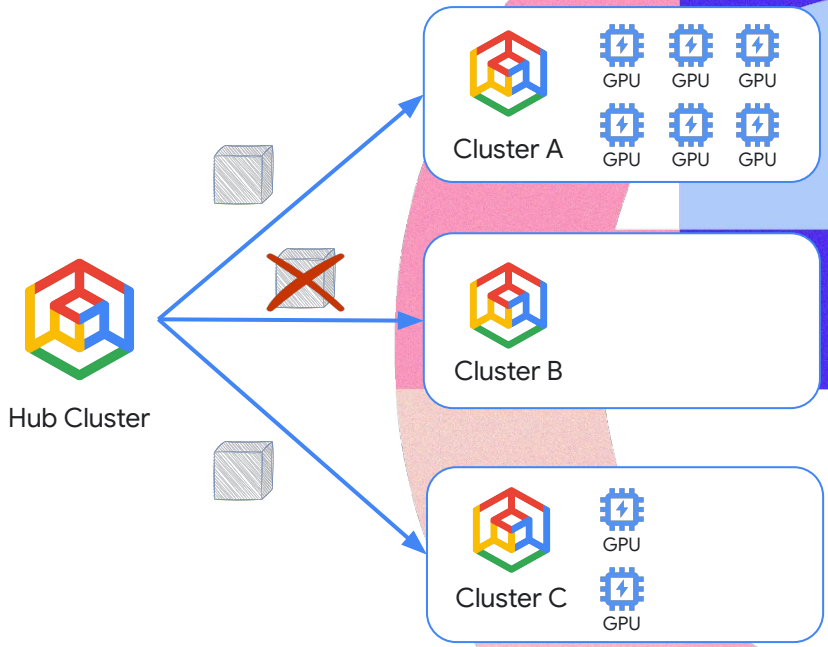
複数クラスタ間のジョブスケジューリング

今までの課題

リクエストがバーストするなどの状況で単一のリージョンにおける GPU などの計算リソースが枯渇した場合に他のリージョンへのフェール オーバー処理を実装する必要があった

解決策

特定のメトリクスを基にあらかじめ用意している複数のクラスタにジョブをスケジューリング可能とすることで計算リソース不足などの影響が抑止可能



GKE Inference Gateway

推論ワークロードに最適化されたインテリジェントな負荷分散

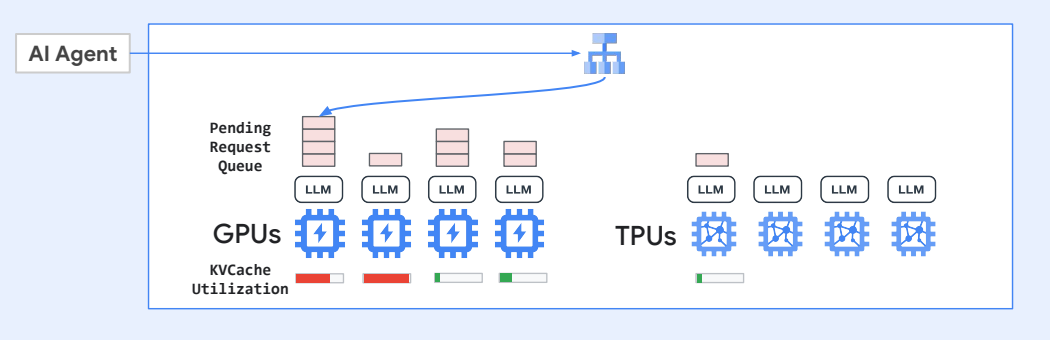
今までの課題

推論処理はリクエストの内容によって計算負荷や必要なメモリ量が大きく変動するため、従来のラウンドロビン方式などの負荷分散では特定のサーバーに負荷が集中してリソース使用率の偏りやリクエストがキューで待たされる事によるコスト効率の悪化、レスポンス遅延が起こり得る

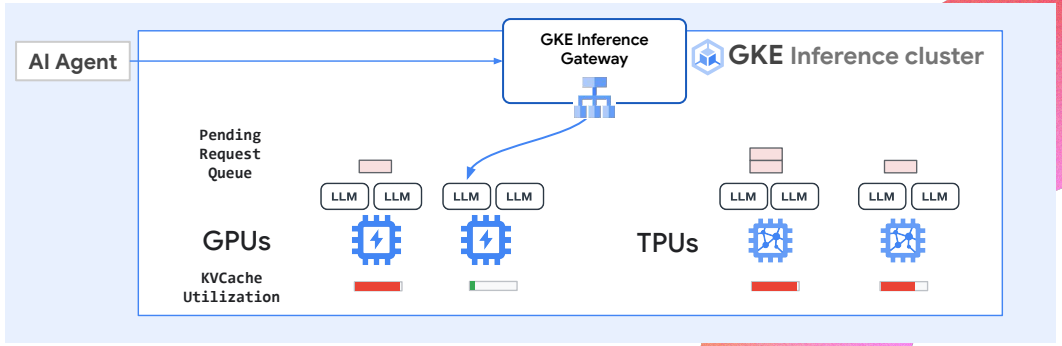
解決策

各サーバーの KVCache 使用率や保留中のリクエスト キューの長さなどのメトリクスを監視し、リクエストをその時点で最も処理能力に余裕のある最適なサーバーへ動的にルーティング

通常のロードバランサー

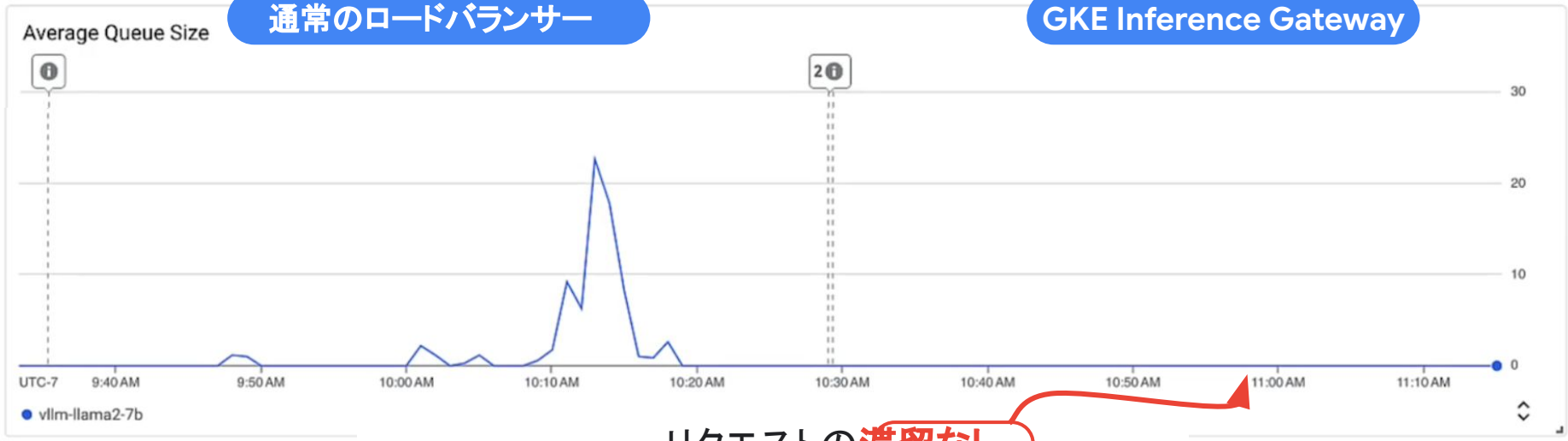


GKE Inference Gateway



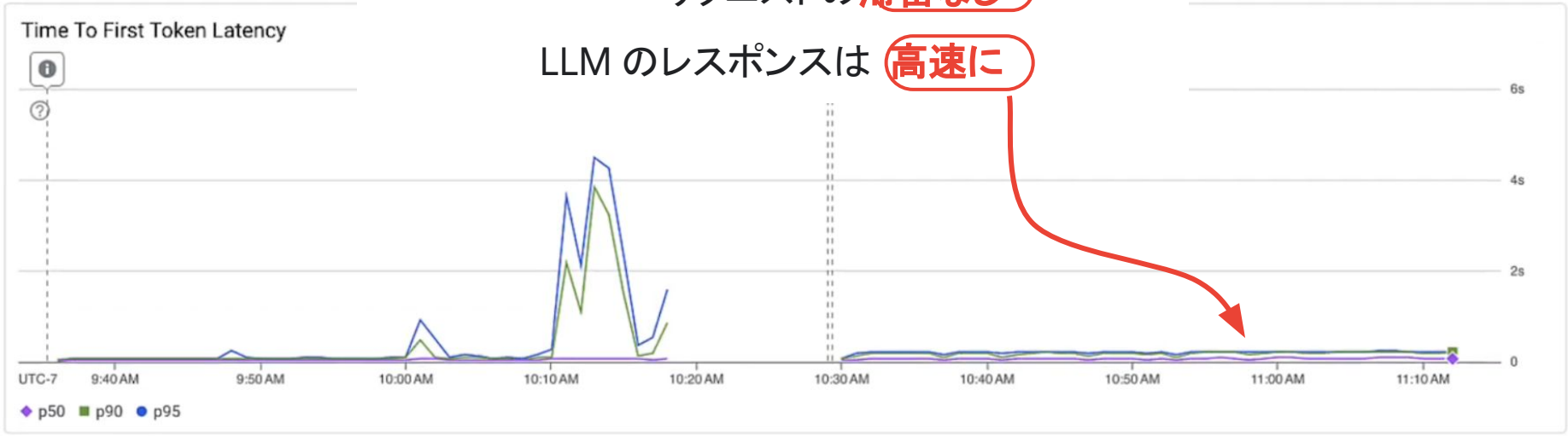
通常のロードバランサー

GKE Inference Gateway



リクエストの滞留なし

LLM のレスポンスは 高速に



Google Kubernetes Engines for AI/ML
さまざまなモデルを簡単にデプロイ

“スタート地点” の提供

コンソールから簡単にモデルがデプロイでき、また Google Cloud が用意した推論用のコンテナ (vLLM ベース) を利用してすぐに GKE 上でモデルがホスティング可能

管理コストの削減

各モデルは GKE の Autopilot クラスタへデプロイが可能なため、ノード自体の管理をせずに簡単にモデルが利用可能

エンタープライズ レディー

安定した運用のためのオブザーバビリティやセキュリティ機能、また高いスケーラビリティによってエンタープライズ アプリケーションにも安心して利用可能

Managed Lustre

高速、かつフルマネージドの分散ファイルシステム

使い慣れたファイル システム

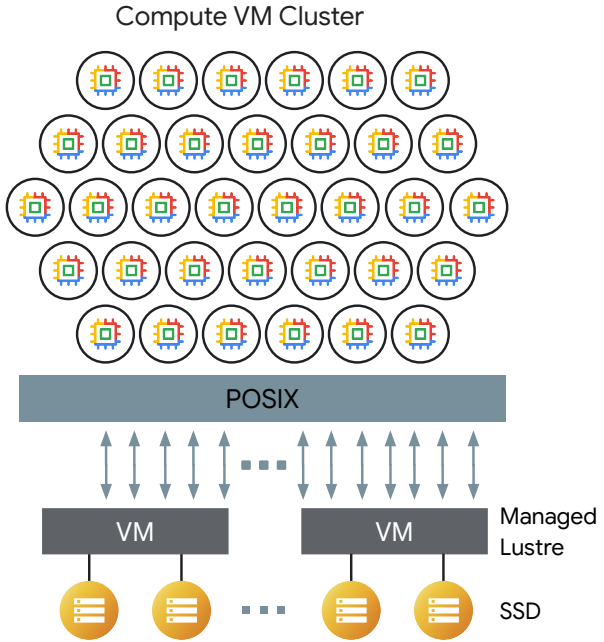
多くのユーザーが利用している Lustre のマネージド サービスであるため、導入のハードルが低い (DDN 社とのパートナーシップ により DDN EXAScaler を利用)

極めて低いレイテンシ

ワークロードに近い VM 上で稼働、また POSIX API 完全互換のインターフェースでのアクセスとなるために極めて低いレイテンシを実現

高い管理性とスケーラビリティ

最大 1PB までのデータが保存可能、またマネージド サービスであるために大規模な分散ファイル システムをシンプルに管理可能



02. AI インフラ運用の難しさ

AI インフラの運用は結構大変

Access

GPU の取得が困難

AI ワークロードを実行したいときに適切な GPU をすぐに利用したい

Usability

AI インフラは難しい

AI 関連の技術は難易度が高いために AI インフラ専任のエンジニアが必要

Power

最もパワフルな GPU

Blackwell GPU をはじめとしてニーズに 合わせて幅広いタイプのGPU が必要

ROI

実行効率の最大化

AI 投資には多くのコストが必要なために実行効率の最大化が必要

AI インフラの運用は結構大変

Access

GPU の取得が困難

AI ワークロードを実行したいときに適切な GPU をすぐに利用したい

Usability

AI インフラは難しい

AI 関連の技術は難易度が高いために AI インフラ専任のエンジニアが必要

Power

最もパワフルな GPU

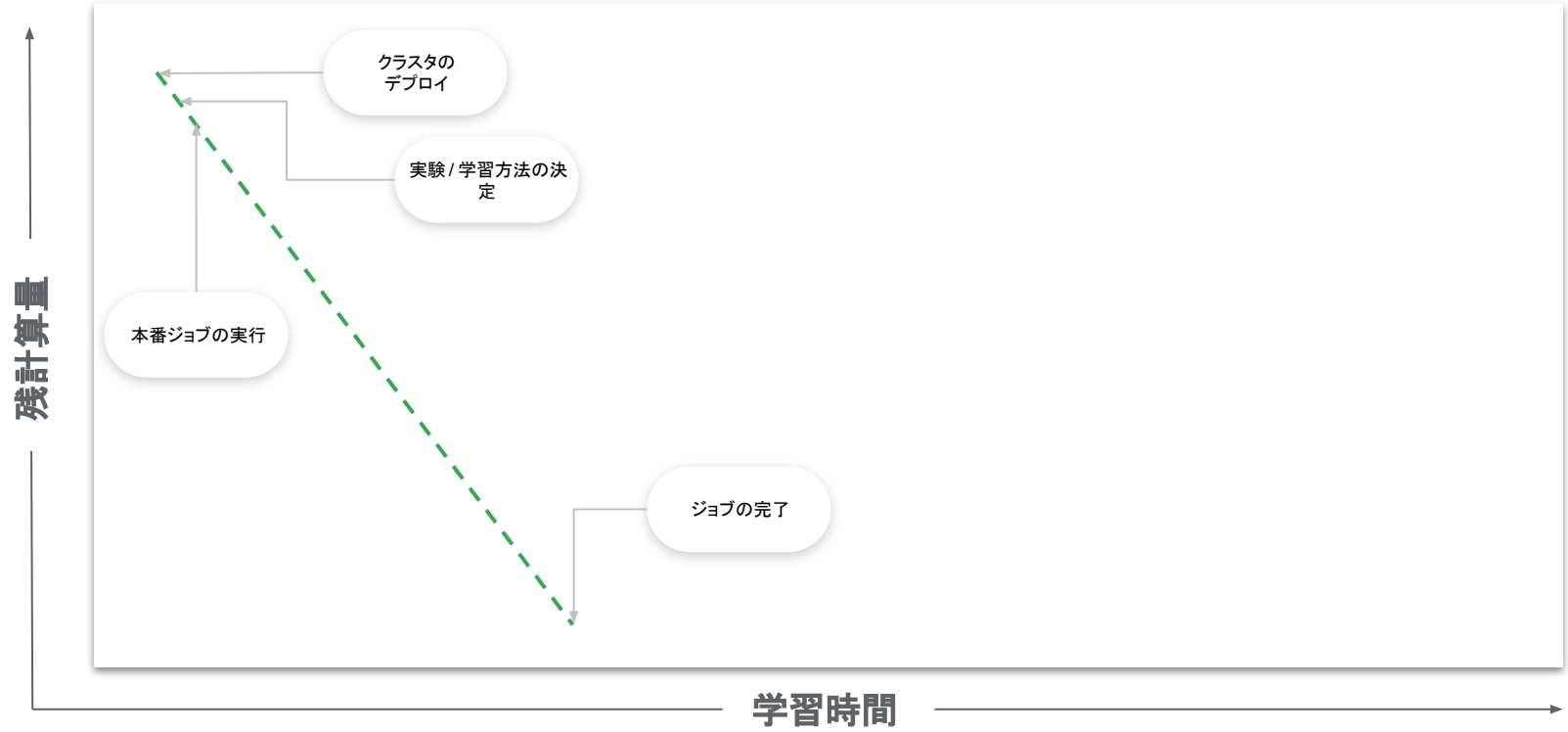
Blackwell GPU をはじめとしてニーズに 合わせて幅広いタイプのGPU が必要

ROI

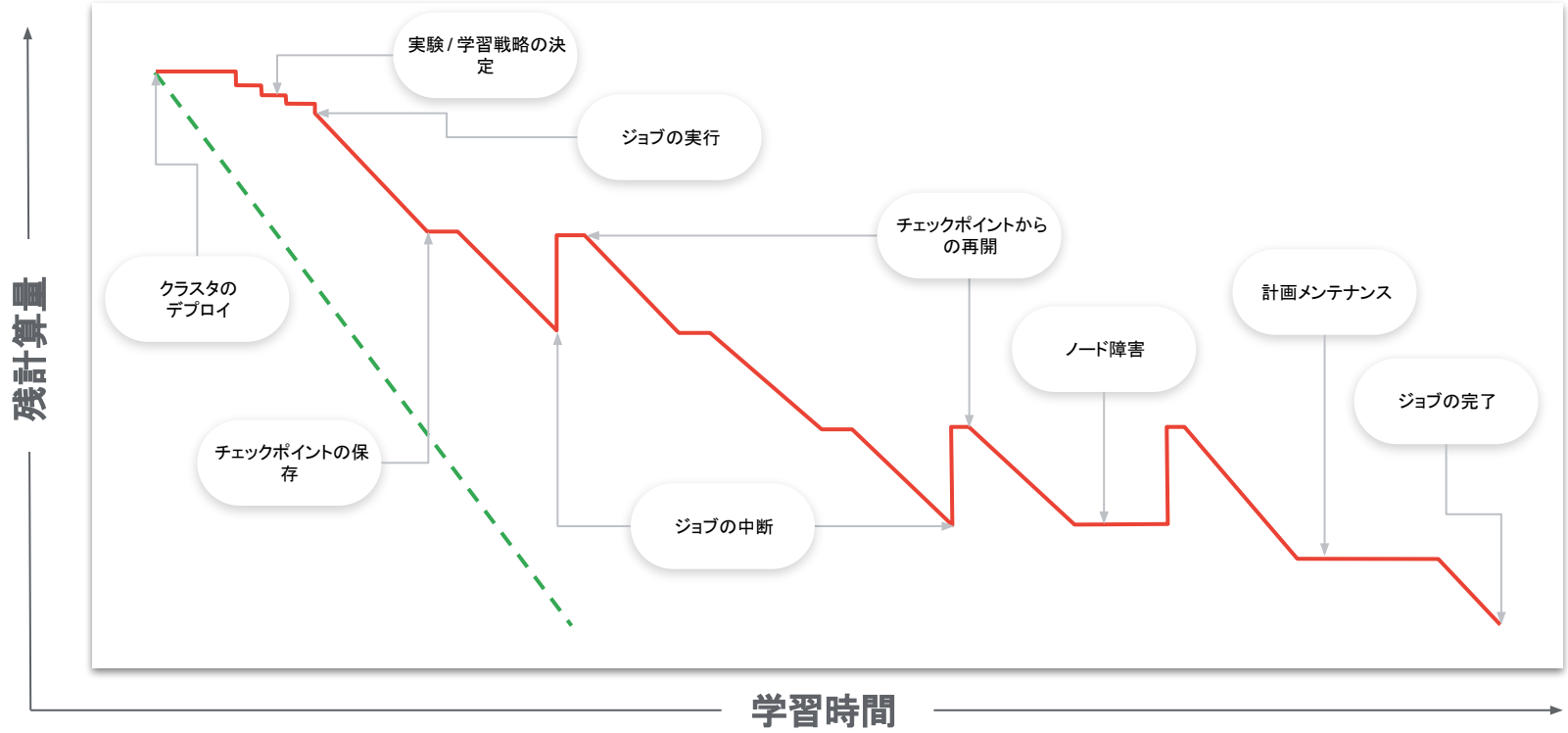
実行効率の最大化

AI 投資には多くのコストが必要なために実行効率の最大化が必要

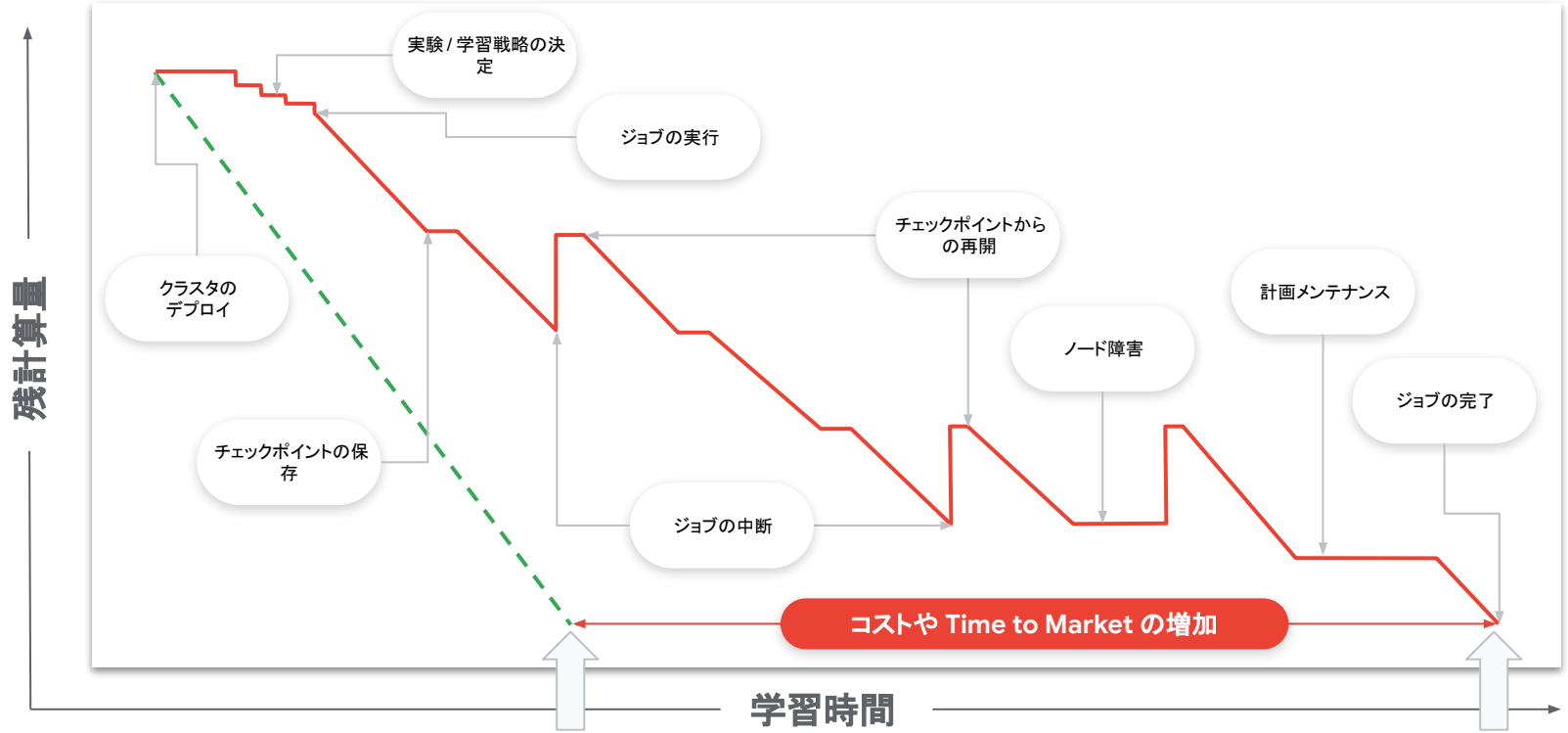
理論的な学習スケジュール



実際の学習スケジュール



実際の学習スケジュール



学習ワークロードの大変さ

01

低いオブザーバビリティ

局所的な障害の検知が極めて困難

02

遅いチェックポイント保存

チェックポイントの保存によるワークロード中断

チェックポイント保存の頻度が低いことによるリストア時の損失

03

さまざまな理由による中断

ワークロードは最新のチェックポイントからの再開が必要

学習生産性の低下

学習効率を測る指標: Goodput

理論的な学習時間と実際の学習時間の差分から計算される Goodput という指標によって E2E での学習効率を計測

$$\text{Goodput} = \frac{\text{IdealChipHours}}{\text{PracticalChipHours}}$$

Goodput の向上への考慮事項	潜在的な Goodput への影響
ハードウェア障害やシステムエラー	5 ~ 15%
Preemption (Spot) やEviction (DWS)	5 ~ 10%
チェックポイント保存やロード時間の遅延	3 ~ 10%
最適ではないチェックポイントの保存頻度	2 ~ 8%
Straggler やパフォーマンスのボトルネック	3 ~ 7%
障害検知時間や素早い診断	2 ~ 5%

https://github.com/AI-Hypercomputer/gpu-recipes/blob/main/training/a3mega/llama3-1-70b/nemo-pretraining-gke-resiliency/goodput-guide.md

学習効率を測る指標: Goodput

理論的な学習時間と実際の学習時間の
差分から計算さ
標によって E2E

Goodput =
PracticalChipHours

さまざまな理由による中断

遅いチェックポイント保存

低いオブザーバビリティ

Goodput の向上への考慮事項	潜在的な Goodputへの影響
ハードウェア障害やシステムエラー	5 ~ 15%
Preemption (Spot) やEviction (DWS)	5 ~ 10%
チェックポイント保存やロード時間の遅延	3 ~ 10%
最適ではないチェックポイントの保存頻度	2 ~ 8%
Straggler やパフォーマンスのボトルネック	3 ~ 7%
障害検知時間や素早い診断	2 ~ 5%

https://github.com/AI-Hypercomputer/
-pretraining-gke-resiliency/goodput-guide.md

Goodput を高める

学習ワークロードの大変さ

01

低いオブザーバビリティ

局所的な障害の検知が極めて困難

02

遅いチェックポイント保存

チェックポイントの保存によるワークロード中断

チェックポイント保存の頻度が低いことによるリストア時の損失

03

さまざまな理由による中断

ワークロードは最新のチェックポイントからの再開が必要

学習生産性の低下

Cluster Director

インフラのコントロール

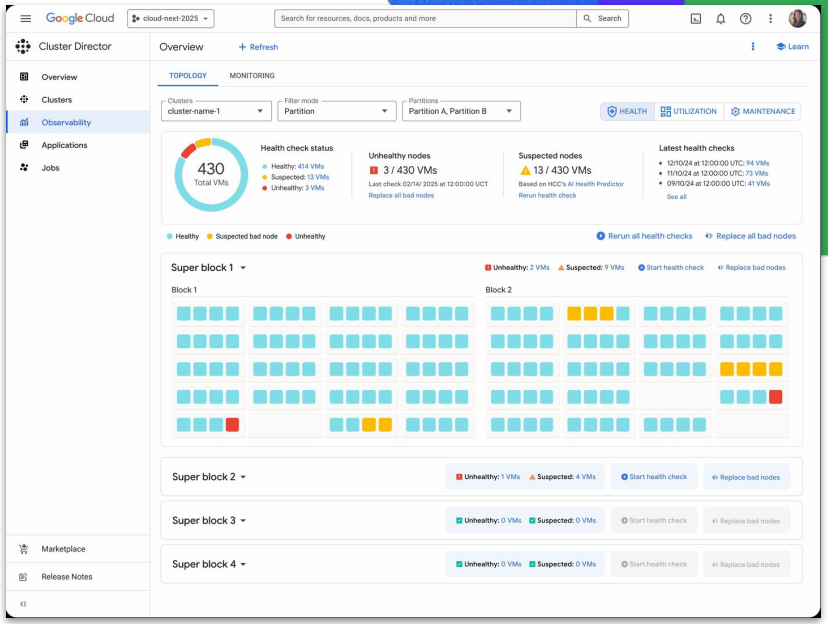
VM を物理的に近い距離にデプロイする、メンテナンス時間を制御する (既定で 90 日周期) といったインフラのコントロールが可能

可観測性と信頼性

AI によるノード障害予測およびヘルスチェック、障害ノードのリプレイス、ネットワークを含むメトリック確認、各ジョブのパフォーマンス確認機能などを提供

インテグレーションとツール

上記の機能は GKE や Slurm に組み込まれ、Cluster Toolkit で簡単にクラスタのデプロイが可能 (将来的には UI でクラスタがデプロイ可能となる予定)



※ UI は今後実装予定です

すでに多くの機能は利用可能

Cluster Health Scanner

GPU のヘルスチェック (DCGM) や NCCL を利用した通信の正常性確認など、複数の構成要素の正常性確認が可能

CoMMA

Collective Communication Analyzer (CoMMA) によって NCCL の Profiler API を利用してテレメトリ データを取得

Report Faulty Host API

CHS や CoMMA などによって正常状態が確認できないノードの交換対応をユーザーが主導で実施するための API

学習ワークロードの大変さ

01

低いオブザーバビリティ

局所的な障害の検知が極めて困難

02

遅いチェックポイント保存

チェックポイントの保存によるワークロード中断

チェックポイント保存の頻度が低いことによるリストア時の損失

03

さまざまな理由による中断

ワークロードは最新のチェックポイントからの再開が必要

学習生産性の低下

Google Cloud Resiliency Framework

分散学習の Goodput が向上

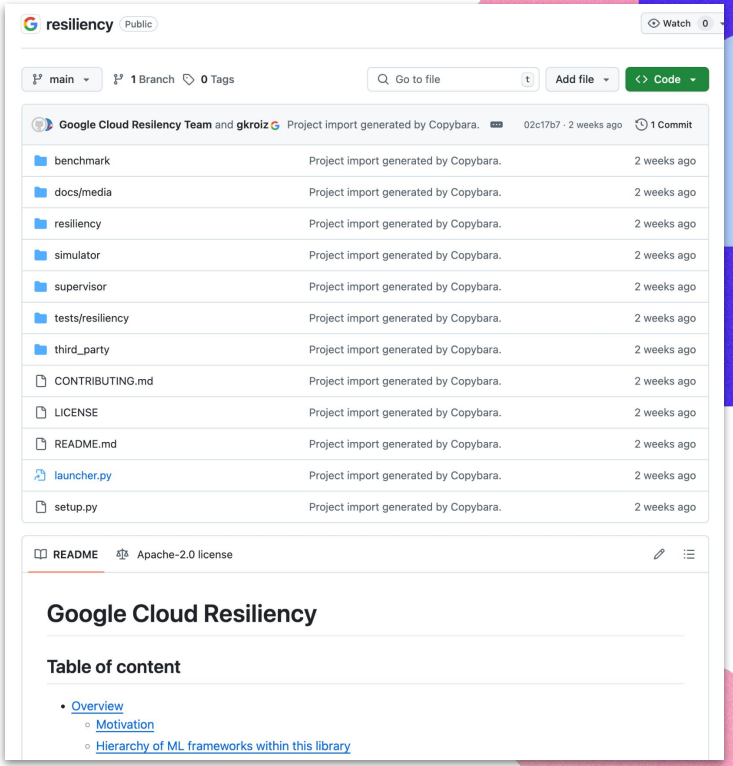
大規模クラスタでの分散学習時における課題を解決することを目指し、ノード障害発生時のスムーズなジョブ再開などを実現するためのフレームワーク

NVIDIA NVRx Library がベース

NVIDIA 社の Resiliency Extension をベースに実装されており、NeMo Framework で実装された学習ジョブと高い親和性

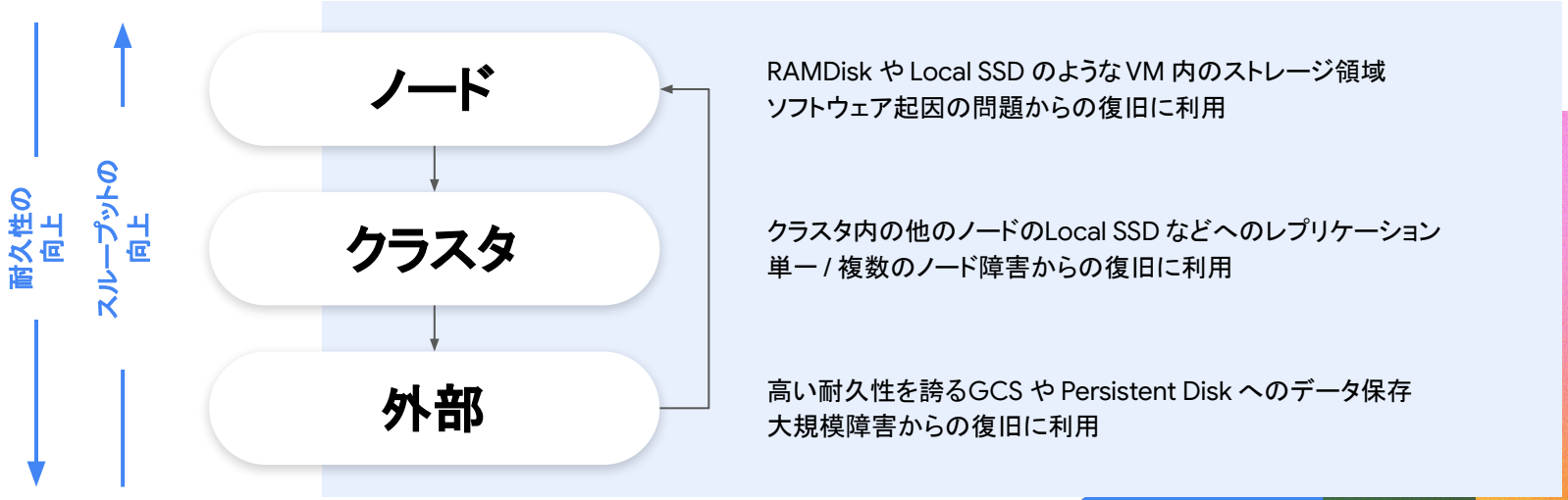
サンプルのご提供

Resiliency Framework が簡単に利用可能なコンテナ イメージや NVIDIA NeMo Framework を利用したサンプルを提供



Multi Tier Checkpointing

各フォールト ドメインのストレージ層：



各層でのデータ レプリケーションは自動的に実施

Multi Tier Checkpointing のメリット

非同期チェックポイント

チェックポイントの取得処理を学習プロセスとは非同期に実行することで GPU がブロッキングされる時間が大幅に減少

高速な復旧

ローカル SSD や同一クラスタ内のノード上のストレージなど、一般的なストレージ サービスと比較し高速なストレージを利用することによる高速は復旧

ロストステップの減少

ローカル SSD への保存や非同期チェックポイントなどの方法によってチェックポイントの取得が高速化することでより高頻度なチェックポイント保存が可能

学習ワークロードの大変さ

01

低いオブザーバビリティ

局所的な障害の検知が極めて困難

02

遅いチェックポイント保存

チェックポイントの保存によるワークロード中断

チェックポイント保存の頻度が低いことによるリストア時の損失

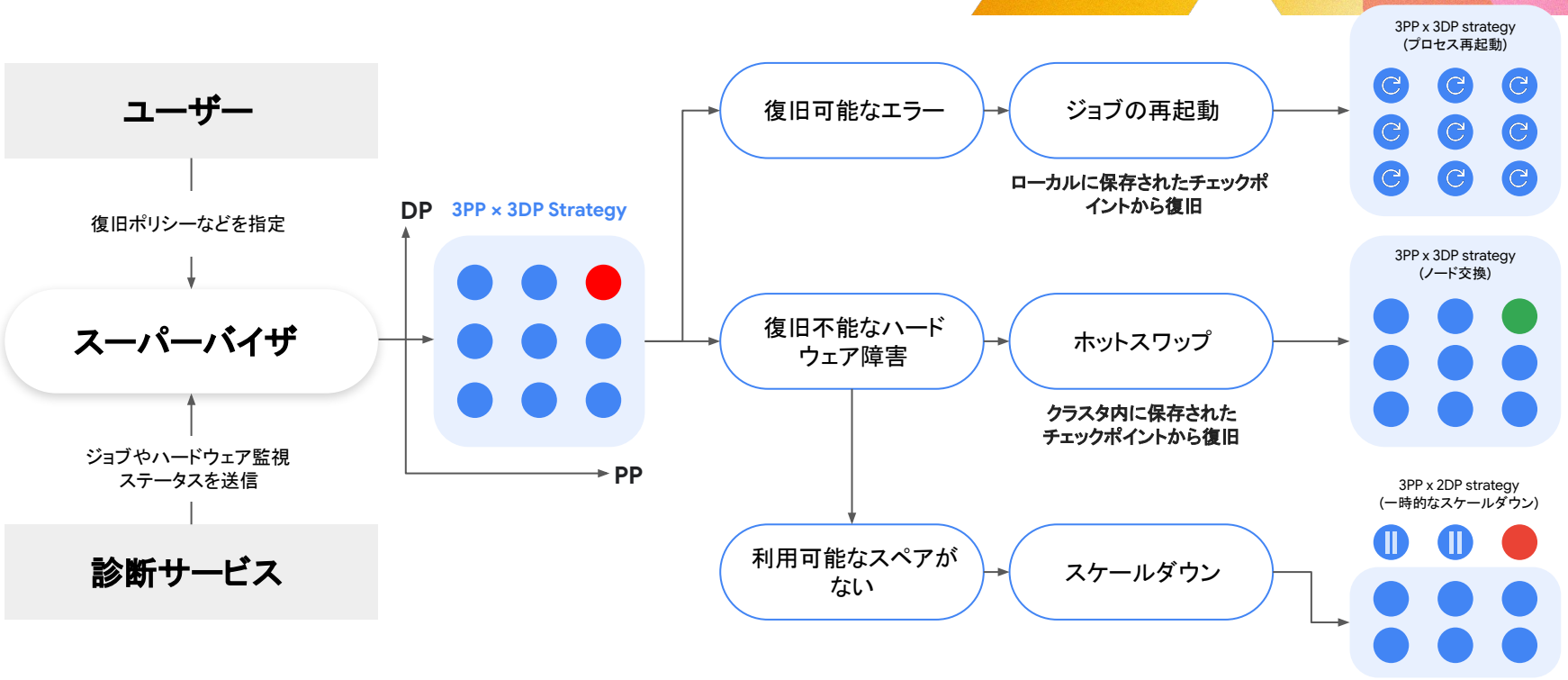
03

さまざまな理由による中断

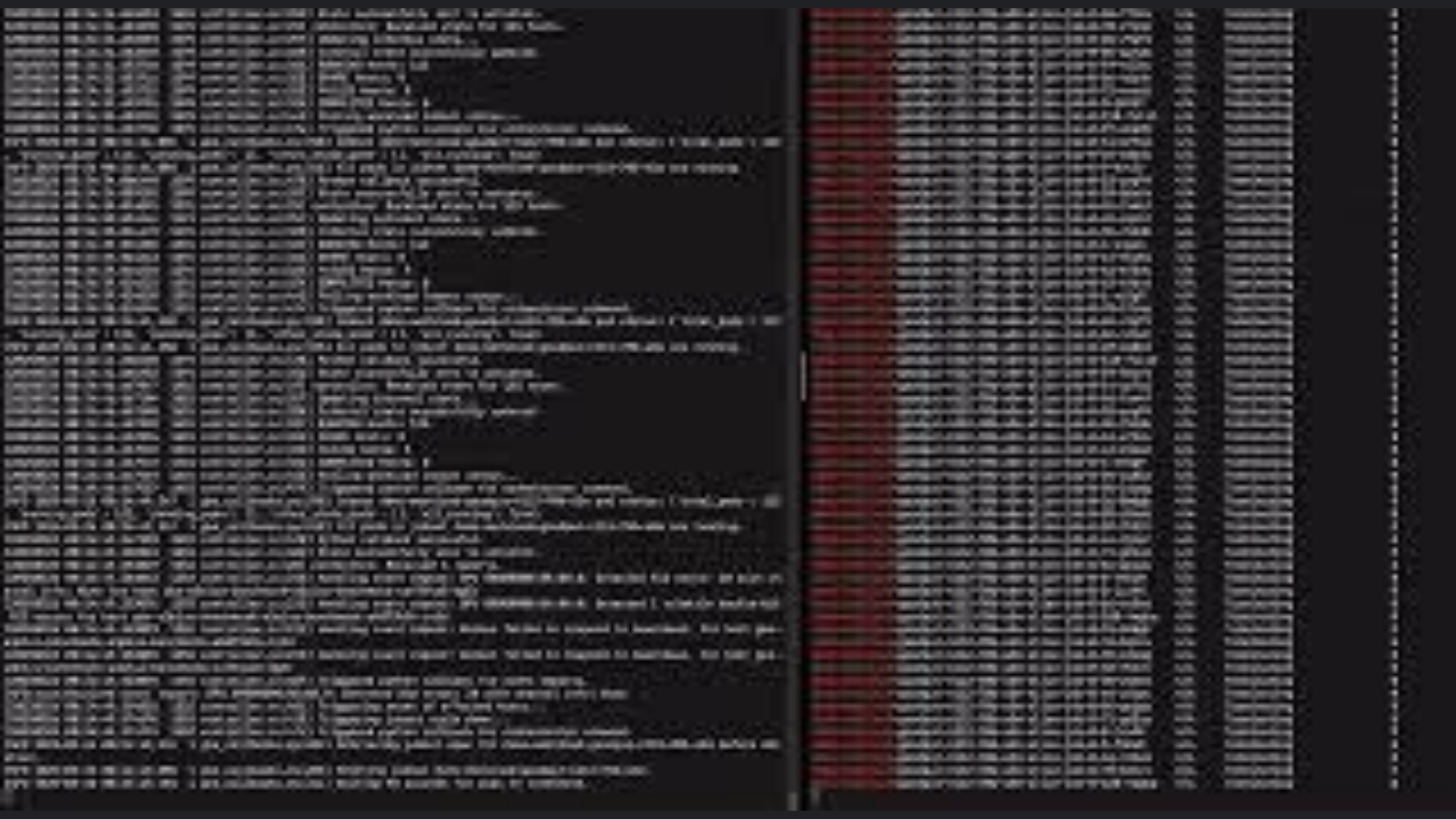
ワークロードは最新のチェックポイントからの再開が必要

学習生産性の低下

継続的なジョブ実行


ユーザー
復旧ポリシーなどを指定
スーパーバイザ
ジョブやハードウェア監視
ステータスを送信
診断サービス
DP
3PP × 3DP Strategy
PP
復旧可能なエラー
ジョブの再起動
ローカルに保存されたチェックポイントから復旧
3PP x 3DP strategy
(プロセス再起動)
復旧不能なハードウェア障害
ホットスワップ
クラスタ内に保存されたチェックポイントから復旧
3PP x 3DP strategy
(ノード交換)
利用可能なスペアがない
スケールダウン
3PP x 2DP strategy
(一時的なスケールダウン)

動画あり
アーカイブ動画をご視聴ください



03. まとめ

AI Hypercomputer

- 最先端のアクセラレータが選択可能
- 高いパフォーマンスとコスト効率
- 大規模クラスタでの信頼性とセキュリティ
- すぐに始められて柔軟にスケール可能